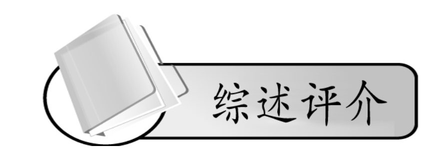


文本相似度计算方法研究综述

陈二静[1,2]　姜恩波[1]

[1](中国科学院成都文献情报中心　成都　610041)

[2](中国科学院大学　北京　100049)

摘要: 【**目的**】分析文本相似度计算方法，了解该领域的发展态势。【**文献范围**】在 CNKI 和 Web of Science 中分别以检索式"篇名: 文本相似度 OR 篇名: 词汇相似度 OR 篇名: 语义相似度"和"TI: 'text similarity' or 'semantic similarity' or 'lexical similarity' "并限定文献类型进行检索，最终得到 69 篇重点文献。【**方法**】对文本相似度计算方法进行系统梳理，分析重点方法的基本思想、特点并总结未来发展方向。【**结果**】形成了较为全面的分类描述体系，文本相似度计算方法可分为 4 类: 基于字符串的方法、基于语料库的方法、基于世界知识的方法和其他方法。其中，基于神经网络和基于世界知识的方法以及针对跨领域文本的相似度计算将成为该领域的发展趋势。【**局限**】仅将不同方法本身作为探讨的核心，未进一步分析方法的应用情况。【**结论**】有助于全面把握和深入了解文本相似度计算方法的研究现状和未来趋势。

关键词: 文本相似度　语义相似度　本体　词袋模型　神经网络

分类号: TP391　G35

1　引　言

在信息爆炸时代，人们迫切希望从海量信息中获取与自身需要和兴趣吻合度高的内容。为了满足此需求，出现了多种应用，如搜索引擎、自动问答系统、文档分类与聚类、文献查重、文献精准推送等，而这些应用场景的关键技术之一就是文本相似度计算技术。近年来，文本相似度受到研究人员的广泛关注，有学者对相关文献进行梳理，总结了文本相似度计算方法[1-2]、词语或词汇相似度算法[3-6]、基于本体的语义相似度算法[7-8]，但有明显不足: 部分文献对国内进展分析较少，未能体现出国内学者在文本相似度方面所取得的进展与成果[1]; 综述局限于文本相似度的某一分支方法，各有侧重点，但覆盖面不全[3-8]; 还有文献将文本相似度分为两类——基于统计或者语料库方法和基于世界知识的文本相似度计算方法，这种分类忽略了基于字符串的方法和句法分析等重要算法[3-5]。随着时间推移，文本相似度计算出现新的研究方法，所以有必要对文本相似度计算方法分类进行扩展。本文旨在对国内外文本相似度计算方法的研究现状进行系统梳理，分析当前各种方法的优缺点，形成较为全面的文本相似度算法分类描述体系，并总结未来发展方向，为相关研究与应用提供参考借鉴。与此同时，本文揭示了文本表示模型的变化以及对文本相似度计算方法的影响。

笔者于 2016 年 11 月 28 日，采用检索式"篇名: 文本相似度 OR 篇名: 词汇相似度 OR 篇名: 语义相似度"在 CNKI 数据库中检索，限制条件为"核心期刊"和收录来源为"CSSCI 中文社会科学引文索引(2016-2017)来源期刊(含扩展版)"，得到中文文献 206 篇; 使用"TI: 'text similarity' or 'semantic similarity' or 'lexical similarity' "的检索式在 Web of Science 核心数据库检索，文献类型为 article，得到 270 篇外文文献。经过清洗、去杂、剔除无效文献，最终筛选出 100 篇文献，笔者在精读的基础上对 69 篇重点文献进行系统梳理。

通讯作者: 陈二静, ORCID: 0000-0002-4663-184X, E-mail: chenerjing@mail.las.ac.cn。

2 文本相似度定义及其相关概念辨析

文本相似度在不同领域被广泛讨论，由于应用场景不同，其内涵有所差异，故没有统一、公认的定义。Lin[9]从信息论的角度阐明相似度与文本之间的共性和差异有关，共性越大、差异越小，则相似度越高；共性越小、差异越大，则相似度越低；相似度最大的情况是文本完全相同。同时基于假设推论出相似度定理，如公式(1)[9]所示。

$$Sim(A,B)=\frac{\log P(common(A,B))}{\log P(description(A,B))} \tag{1}$$

其中，$common(A,B)$ 是 A 和 B 的共性信息，$description(A,B)$ 是描述 A 和 B 的全部信息，公式(1)表达出相似度与文本共性成正相关。由于没有限制应用领域，此定义是被较多采用的概念。

相似度与相关度是容易混淆的概念，大量学者[4,8,10]对此做过对比说明。相关度体现在文本共现或者以任何形式相互关联(包括上下位关系、同义关系、反义关系、部件–整体关系、值–属性关系等[11])，反映出文本的组合特点[12]。而相似度是相关度的一种特殊情况，包括上下位关系和同义关系。由此得出，文本相似度越高，则相关度越大，但是相关度越大并不能说明相似度高。

相似度一般可用[0,1]之间的实数表示，该实数可通过语义距离计算获得。相似度与语义距离呈反比关系，语义距离越小，相似度越高；语义距离越大，相似度越低。通常用公式(2)[10]表示相似度与语义距离的关系。

$$Sim(S_A,S_B)=\frac{\alpha}{Dis(S_A,S_B)+\alpha} \tag{2}$$

其中，$Dis(S_A,S_B)$ 表示文本 S_A、S_B 之间的非负语义距离，α 为调节因子，保证了当语义距离为 0 时公式(2)具有意义。

文本相似度计算中还有一个重要概念是文本表示，代表对文本的基本处理，目的是将半结构化或非结构化的文本转换为计算机可读形式。文本相似度计算方法的不同本质是文本表示方法不同。

3 文本相似度计算方法

大多学者将文本相似度计算方法分为基于统计或者语料库的方法和基于世界知识的方法，这种分类忽略了基于字符串和句法分析等重要算法，且近年来有新的方法出现，所以本文借鉴 Gomaa 等[1]的分类框架，对分类体系进行扩展和细分，如图 1 所示。

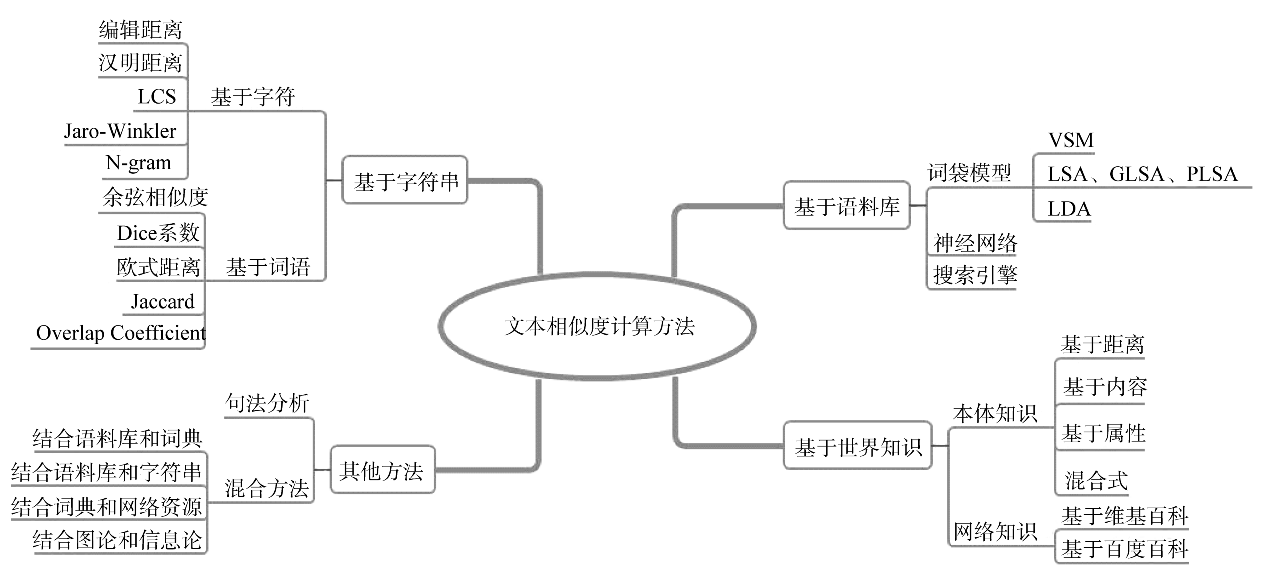


图 1 文本相似度计算方法分类

将文本相似度计算方法分为 4 大类：基于字符串(String-based)的方法、基于语料库(Corpus-based)的方法、基于世界知识(Knowledge-based)的方法和其他方法。基于字符串的方法也称作“字面相似度方法”，其中较为典型的方法包括最长公共子串(Longest Common Substring, LCS)、编辑距离、Jaccard 等。由

于基于字符串的方法没有考虑文本的语义信息，计算效果受到一定限制。为解决这一问题，学者们开始对语义相似度方法展开研究，包括基于字符串的方法、基于语料库的方法、基于世界知识的方法和其他方法。其中其他方法又包括句法分析和混合方法，句法分析是对句子的语法结构分析，也属于语义分析的一种，但其不依赖于某种语料库或世界知识，所以被划分到其他方法。混合方法则是对几种方法的综合。

3.1　基于字符串

该方法从字符串匹配度出发，以字符串共现和重复程度为相似度的衡量标准。根据计算粒度不同，可将方法分为基于字符(Character-Based)的方法和基于词语(Term-Based)的方法。一类方法单纯从字符或词语的组成考虑相似度算法，如编辑距离、汉明距离、余弦相似度、Dice 系数、欧式距离；另一类方法还加入了字符顺序，即字符组成和字符顺序相同是字符串相似的必要条件，如最长公共子串(Longest Common Substring, LCS)、Jaro-Winkler；再一类方法采用集合思想，将字符串看作由词语构成的集合，词语共现可用集合的交集计算，如 N-gram、Jaccard、Overlap Coefficient。表 1 列出了主要方法，其中 S_A、S_B 表示字符串 A、B。

表 1　基于字符串的代表方法

类型	方法	基本思想	类型	特点与不足
基于字符	编辑距离	S_A 转换到 S_B 需要删除、插入、替换操作的最少次数。	字符组成	计算准确，但费时。
	汉明距离[13]	$1-\left(\sum_{k=1}^{n} x_k \oplus y_k\right)/n$，其中 x_k, y_k 分别表示字符串 S_A、S_B 对应码字第 K 位的分量。	字符组成	采用模 2 加运算，简化长文本计算，效率高。
	LCS	共现且最长的子字符串。	字符顺序	原理简单，针对派生词和短文本有较好效果，但不适用于长文本。
	Jaro-Winkler	$d_j=\frac{1}{3}\left(\frac{m}{\|S_A\|}+\frac{m}{\|S_B\|}+\frac{m-t}{m}\right)$，其中 m 是匹配的字符数；t 是换位的数目。相似度计算公式为 $d_j+(lp(1-d_j))$，其中 d_j 是两个字符串的 Jaro 距离，l 是前缀相同的长度，规定最大为 4。Winkler 将 p 定义为 0.1。	字符顺序	考虑了前缀相同的重要性，针对短文本有较好效果，但不适用于长文本。
	N-gram	$\frac{相似的n元组数量}{n元组总量}$	集合思想	n 可调，方法较为灵活，但不适用于长文本。
基于词语	余弦相似度	$\frac{\vec{S_A}\cdot\vec{S_B}}{\|\|S_A\|\|\|\|S_B\|\|}$	词语组成	将文本置于向量空间，解释性强，较为常用，但不适用于长文本。
	Dice 系数[14]	$\frac{2\times comm(S_A,S_B)}{leng(S_A)+leng(S_B)}$	词语组成	增强相同部分的作用，有效关注较短的相同文本。
	欧式距离	$\sqrt{S_A^2+S_B^2}$	词语组成	算法简单直接，但效果粗糙，不适用于长文本。
	Jaccard	$\frac{S_A\cap S_B}{S_A\cup S_B}$	集合思想	不适用于长文本。
	Overlap Coefficient	$\frac{S_A\cap S_B}{\min(S_A,S_B)}$	集合思想	当一个字符串是另一个字符串的子字符串时，相似度最大。

基于字符串的方法是在字面层次上的文本比较，文本表示即为原始文本。该方法原理简单、易于实现，现已成为其他方法的计算基础。但不足的是将字符或词语作为独立的知识单元，并未考虑词语本身的含义和词语之间的关系。以同义词为例，尽管表达不同，但具有相同的含义，而这类词语的相似度依靠基于字符串的方法并不能准确计算。

3.2　基于语料库

基于语料库的方法利用从语料库中获取的信息计算文本相似度。基于语料库的方法可以分为：基于词

袋模型的方法、基于神经网络的方法和基于搜索引擎的方法。前两种以待比较相似度的文档集合为语料库，后一种以 Web 为语料库。

(1) 基于词袋

词袋模型(Bag of Words Model, BOW)建立在分布假说的基础上，即"词语所处的上下文语境相似，其语义则相似"[15]。基本思想是不考虑词语在文档中出现的顺序，将文档表示成一系列词语的组合。根据考虑的语义程度不同，基于词袋模型的方法主要包括向量空间模型(Vector Space Model, VSM)、潜在语义分析(Latent Semantic Analysis, LSA)、概率潜在语义分析(Probabilistic Latent Semantic Analysis, PLSA)和潜在狄利克雷分布(Latent Dirichlet Allocation, LDA)。

①VSM

20 世纪 60 年代末，Salton 等提出 VSM[16]，这种方法受到广大学者的青睐。基本思想是将每篇文档表示成一个基于词频或者词频–逆文档频率(Term Frequency-Inverse Document Frequency, TF-IDF)权重的实值向量，那么 N 篇文档则构成 n 维实值空间，其中空间的每一维都对应词项，每一篇文档表示该空间下的一个点或者向量。而两个文档的相似度就是两个向量的距离，一般采用余弦相似度方法计算。已有学者对 VSM 方法做出改进，如郭庆琳等[17]通过增加关键特征词改进 DF 在特征值选择时过滤有用信息的不足，以及在计算 TF-IDF 时加入特征词筛选阶段的特征权重，从而在没有增加时间和空间复杂度的情况下，提高精确度。李连等[18]针对传统 VSM 算法没有统计文本相同特征词数量而导致计算不准确的问题，引入表征文本特征词覆盖程度的参数，优化了文本相似度的计算结果。

基于 VSM 的方法基本原理简单，但该方法有两个明显缺点：一是该方法基于文本中的特征项进行相似度计算，当特征项较多时，产生的高维稀疏矩阵导致计算效率不高；二是向量空间模型算法的假设是文本中抽取的特征项没有关联，这不符合文本语义表达。

②LSA、PLSA

LSA[19]算法的基本思想是将文本从稀疏的高维词汇空间映射到低维的潜在语义空间，在潜在语义空间计算相似性。LSA 是基于 VSM 提出来的，两种方法都是采用空间向量表示文本，但 LSA 使用潜在语义空间，利用奇异值分解(Singular Value Decomposition, SVD)技术对高维的词条–文档矩阵进行处理，去除了原始向量空间的某些"噪音"，使数据不再稀疏。Hofmann[20]在 LSA 基础上引入主题层，采用期望最大化算法(Expectation Maximization, EM)训练主题，得到改进的 PLSA 算法。LSA 本质上是通过降维提高计算准确度，但该算法复杂度比较高，可移植性差。比较之下，PLSA 具备统计基础，多义词和同义词在 PLSA 中分别被训练到不同的主题和相同的主题下，从而避免了多义词、同义词的影响，使得计算结果更加准确，但不适用于大规模文本。

③LDA

LDA[21]主题模型是一个三层贝叶斯概率模型，包含词、主题和文档三层结构。采用 LDA 计算文本相似性的基本思想是对文本进行主题建模，并在主题对应的词语分布中遍历抽取文本中的词语，得到文本的主题分布，通过此分布计算文本相似度[22]。与 PLAS 不同的是，LDA 的文档到主题服从 Dirichlet 分布，主题到词服从多项式分布，此方法适用于大规模文本集，也更具有鲁棒性。熊大平等[23]提出利用 LDA 计算问句相似度，将查询语句和问题分别用 LDA 主题分布概率表示，采用余弦相似度计算二者的相似度，效果有了一定的提高，尤其对特征词不同但主题相似的问题有突出效果，该方法适用于单个问句。张超等[24]将 LDA 分别应用于文本的名词、动词和其他词，得到不同词性词语的相似度，综合加权三个相似度计算文本相似度，此方法由于将建模过程并行化，从而降低了时间复杂度。

以上三类尽管都是采用词袋模型实现文本表示，但是不同方法考虑的语义程度有所不同。基于向量空间模型的方法语义程度最低，仅仅建立在分布假说理论基础上，而忽略了词语之间的关联。基于 LSA、PLSA 的方法语义程度居中，加入潜在语义空间概念，解决了向量空间模型方法的稀疏矩阵问题并降低了多义词、同义词的影响。基于 LDA 主题模型的方法语义程度最高，基于相似词语可能属于同一主题的理论，主题经过训练得到，从而保证了文本的语义性。

(2) 基于神经网络

通过神经网络模型生成词向量(Word Vector、Word Embeddings 或 Distributed Representation)[25-26]计算文本相似度是近年来自然语言处理领域研究较多的方法。不少产生词向量的模型和工具也被提出，如 Word2Vec[27]和 GloVe[28]等。词向量的本质是从未标记的非结构文本中训练出的一种低维实数向量，这种表达方式使得类似的词语在距离上更为接近，同时较好地解决了词袋模型由于词语独立带来的维数灾难和语义不足问题。Kenter 等[29]合并由不同算法、语料库、参数设置得到的不同维度词向量并训练出特征，经过监督学习算法得到训练分类器，利用此分类器计算未标记短文本之间的相似度分数。Kusner 等[30]提出使用词向量计算文档相似度的新方法，即在词向量空间里计算将文档中所有的词移动到另一文档对应的词需要的最小移动距离(Word Mover's Distance, WMD)，求解出来的 WMD 则是两个文档的相似度。Huang 等[31]在

WMD 的基础上提出改进方法——监督词移动距离(Supervised-WMD, S-WMD), 实质上加入新文档特征"re-weighting"和新移动代价"metric A", 令 WMD 方法适用于可监督的文本。

基于神经网络方法与词袋模型方法的不同之处在于表达文本的方式。词向量是经过训练得到的低维实数向量, 维数可以人为限制, 实数值可根据文本距离调整, 这种文本表示符合人理解文本的方式, 所以基于词向量判断文本相似度的效果有进一步研究空间。

(3) 基于搜索引擎

随着 Web3.0 时代的到来, Web 成为内容最丰富、数据量最大的语料库, 与此同时搜索引擎相关算法的进步, 使得有任何需求的用户都可通过搜索找到答案。自从 Cilibrasi 等[32]提出归一化谷歌距离(Normalized Google Distance, NGD)之后, 基于搜索引擎计算语义相似度的方法开始流行起来。其基本原理是给定搜索关键词 x、y, 搜索引擎返回包含 x、y 的网页数量 $f(x)$、$f(y)$以及同时包含 x 和 y 的网页数量 $f(x, y)$, 计算谷歌相似度距离如公式(3)[32]所示。

$$NGD(x,y)=\frac{G(x,y)-\min(G(x),G(y))}{\max(G(x),G(y))}=\frac{\max\{\log f(x),\log f(y)\}-\log f(x,y)}{\log N-\min\{\log f(x),\log f(y)\}} \quad (3)$$

但是该方法最大的不足是计算结果完全取决于搜索引擎的查询效果, 相似度因搜索引擎而异。刘胜久等[33]采用多个搜索引擎的搜索结果, 根据搜索引擎的市场份额为其赋予权重, 得到的结果更加综合全面。此方法简单, 避免了单个搜索引擎所导致的偏差, 但是没有对各搜索结果进行重要性分析。一些学者提出通过分析返回网页内容计算相似度, Sahami 等[34]将查询关键词返回的网页内容构建为语境向量(Context Vector), 采用相似度核函数计算语境向量之间的相似度, 比单纯使用搜索数量计算相似度有更丰富的语义信息。第三类方法是综合搜索结果数量和搜索结果内容, 陈海燕[35]定义了语义片段, 即两个关键词共同出现的片段, 通过分析网页内容获取语义片段数量, 替换包含两个关键词的网页数量, 得到较为精确的相似度。

基于搜索引擎的相似度方法为相似度计算提供了丰富的语义信息, 计算结果依赖于搜索引擎的搜索效果以及对网页内容的语义分析效果, 所以精确获取返回网页数量和有效分析网页内容成为关键问题。

3.3 基于世界知识

基于世界知识的方法是指利用具有规范组织体系的知识库计算文本相似度, 一般分为两种: 基于本体知识和基于网络知识。前者一般是利用本体结构体系中概念之间的上下位和同位关系, 如果概念之间是语义相似的, 那么两个概念之间有且仅有一条路径[7,10]。而网络知识中词条呈结构化并词条之间通过超链接形式展现上下位关系, 这种信息组织方式更接近计算机的理解。概念之间的路径或词条之间的链接就成为文本相似度计算的基础。

(1) 基于本体

文本相似度计算方法使用的本体不是严格的本体概念, 而指广泛的词典、叙词表、词汇表以及狭义的本体。随着 Berners-Lee 等提出语义网的概念, 本体成为语义网中对知识建模的主要方式, 在其中发挥着重要作用。由于本体能够准确地表示概念含义并能反映出概念之间的关系, 所以本体成为文本相似度的研究基础[7]。最常利用的本体是通用词典, 例如 WordNet、《知网》(HowNet)和《同义词词林》等, 除了词典还有一些领域本体, 例如医疗本体、电子商务本体、地理本体、农业本体等。

结合 Hliaoutakis[36] 、Batet 等[37]的研究, 将基于本体的文本相似度算法概括为 4 种: 基于距离(Edge-Counting Measures)、基于内容(Information Content Measures)、基于属性(Feature-based Measures)和混合式(Hybrid Measures)相似度算法。表 2 列出了各种方法的基本原理、代表方法和特点。

基于本体的方法将文本表示为本体概念以及概念之间的关系, 该方法能够准确反映概念内在语义关系, 是一种重要的语义相似度计算方法, 主要缺点如下:

①本体一般需要专家参与建设, 耗费大量时间和精力, 而已有的通用本体存在更新速度慢、词汇量有限等问题, 不适用于出现的新型词语;

②利用本体计算文本相似度, 首先是在词语层次进行计算, 然后累加词语相似度获得长文本相似度, 相对基于语料库的方法对文本整体处理而言计算效率较低;

③无论是通用本体还是领域本体, 本体之间相互独立将带来本体异构问题, 不利于跨领域的文本相似度计算。

表 2　基于本体的方法

	基于距离	基于内容	基于属性	混合式
基本原理	用概念之间的路径长度表示语义距离	用概念词共享的信息量化它们之间的语义相似度	用概念词之间的公共属性数量衡量它们之间的相似度	将基于距离、基于内容、基于属性三种方法综合计算概念之间的相似度
代表方法	Shortest Path[38]、Wu 等[39]、Weighted Links[40]、Li 等[41]、刘群等[10]	Lin[42]、Resnik[43]、Lord 等[44]、边振兴[45]	Tversky[46]	葛斌等[47]、王艳娜等[48]、李文清等[49]
特点	在计算方法中加入了节点深度、密度、强度、宽度及分类体系层次等影响因子	计算方法采用不同节点的信息量以及表达信息内容的不同公式	计算效果依赖于本体属性集的完整性	计算方法中权重参数设置大多依赖领域专家

(2) 基于网络知识

由于本体中词语数量的限制，有些学者开始转向基于网络知识方法的研究，原因是后者覆盖范围广泛、富含丰富的语义信息、更新速度相对较快，使用最多的网络知识是维基百科、百度百科。网络知识一般包括两种结构，分别是词条页面之间的链接和词条之间的层次结构。孙琛琛等[50]将其概括为：文章网络和分类树(以树为主题的图)。

最早使用维基百科计算语义相关度是 Strube 等[51]提出的 WikiRelate!方法，基本原理是在维基百科中检索出与词语相关的网页，并通过抽取网页所属类别找到分类树，最终基于抽取的页面以及在分类法中的路径计算相关度。该方法利用了维基百科的层次结构，计算效果与基于本体的方法相当，然而此方法更适用于词语丰富的文本。Gabrilovich 等[52]提出 ESA 方法，基于维基百科派生出高维概念空间并将词语表示为维基百科概念的权重向量，通过比较两个概念向量(比如采用余弦值方法)得到语义相关度，计算效果优于人工判读。ESA 比 WikiRelate!表达更加复杂的语义，而且模型对用户来说简单易懂，鲁棒性较好。Milne 等[53]提出的 WLM 方法仅使用维基百科的链接结构以及较少的数据和资源，比 ESA 简单，但计算结果不如 ESA 理想。严格来说，这些方法是计算文本语义相关度，其包括范围比语义相似度大，但是这些方法为基于维基百科的语义相似度计算提供了良好的借鉴。盛志超等[54]提出一种模仿人脑联想方式的方法，基于维基百科页面的链接信息，并依托 TF-IDF 算法得到词语相似度，尽管取得了一定的效果，但是将维基百科的页面信息和类别信息以较为简单的方式结合成统一的知识源，过于简单，缺乏一定的理论支撑。彭丽针等[55]考虑到维基百科页面的社区现象[56]，对带有标签的页面采用 HITS 算法获取社区类别，基于词语类别与链接关系计算相似度，实验证明该方法具有一定的可行性和有效性，但由于未深入分析页面内容导致语义程度较弱。

与维基百科类似，百度百科作为众人参与可协作的中文百科全书，到 2017 年 1 月已经有超过 1 400 万的词条，数据量成为百度百科相较于其他语料库的绝对优势。詹志建等[57]在分析百科词条结构的基础上，采用向量空间模型计算百科名片、词条正文、相关词条的相似度，采用基于信息内容的方法计算开放分类的相似度，最终加权得到词条相似度，计算效果优良，但是该方法对词条语义信息的分析并不深入。尹坤等[58]在计算方法中引入图论思想，将百度百科视为图，词条视为图中节点，采用 SimRank 方法计算词条之间的相似度。该方法充分利用了百科词条之间的链接关系，但仅对于相关词条较多的词条有好的效果，而对于相关词条较少的词条的计算效果则不理想。

综上所述，基于网络知识的文本相似度计算方法大多利用页面链接或层次结构，能较好地反映出词条的语义关系。但其不足在于：词条与词条的信息完备程度差异较大，不能保证计算准确度；网络知识的产生方式是大众参与，导致文本缺少一定的专业性。

3.4　其他方法

除了基于字符串、基于语料库和基于世界知识的方法，文本相似度计算还有一些其他方法，本文将研究较多的句法分析和混合方法作为其他方法的代表进行具体阐述。

(1) 句法分析

文本相似度方法一般以词语为粒度，而较少关注

词语的组合方式和组合内涵，也就是句法分析。句法是文本语句的重要组成部分，相同词语经由不同句法组织之后所表达的含义差别很大，所以句法分析对计算句子粒度的相似度有着重要作用。

穗志方等[59]提出"骨架依存分析法"并基于此方法设计语句相似度计算模型，基本思想是分析句子的谓语中心词以及其直接支配成分，将分析结果以依存树的形式表达出来，通过比较骨架依存树得到文本相似度。该方法给出单句相似度计算方法，适用于问答系统应用场景，但针对全文计算相似度时，要依次分析语句成分并构建依存树就会造成巨大工作量，所以该方法不适用于长文本。李彬等[60]仅考虑为有效搭配对构建依存树，即句子中的动词、名词和形容词及其直接支配成分，大大降低了计算复杂度和时间成本，但对于包含较多动词的长句效果不好。李茹等[61]基于汉语框架网(CFN)类语义资源，采用多框架描述句子，通过比较重要度高的框架计算句子相似度。Blanco 等[62]提出三层逻辑形式转换(LFT)的新型句法分析，结合从逻辑验证派生出的语义特征和监督机器学习框架，获得相似度分数，该方法首次完成从句子中抽取语义关系并应用到文本相似度计算中。

基于句法分析的关键是找到句子中各部分的依存关系或语义关系，在计算相似度的同时考虑词语相似度和关系相似度，故此方法具有更丰富的语义，但是句子本身的复杂性为框架分析带来的难度和工作量不容小觑，目前研究基本从两个方面进行改进，有效提取关键词和选择合适的语义框架。

(2) 混合方法

由于单一算法具有一定优势与不足，所以学者综合运用两种或两种以上的方法计算文本相似度。较早时期，Jiang 等[63]将 WordNet 词典分类结构与语料库统计信息结合，通过计算概念相关性判断文本的相关性。Islam 等[64]结合语料库和字符串方法，使用词汇数量级在 10^8 的语料库，对于较短字符串使用 LCS 方法，既提高了计算效果，又降低了时间复杂度。Tasi 等[65]将 VSM 和 LCS 结合起来，同时考虑文本的序列关系和权重，有效地提高了准确率。魏韡等[66]结合图论与信息量理论提出一种混合方法，同时考虑词语所在的有向无环图和处于不同位置的节点内在信息量，计算结果比较符合人工判断。Liu 等[67]结合 WordNet 的结构表达信息和网络资源的统计信息计算文本相似度并取得了较好的效果。王小林等[68]在 TF-IDF 算法的基础上加入信息熵和信息增益并结合语义加权因子，最终得到的文本相似度更接近现实。Atoum 等[69]利用基于距离和基于信息内容方法计算词语相似度，将词语相似度进行加权并融合句子长度得到文本相似度。

混合方法是学者对不同方法结合方式的探索，在一定程度上提高文本相似度计算效果。由于文本相似度计算领域的方法颇为丰富，每类方法中的影响因素并不单一，所以混合方法的思路较为开阔，但不可避免的是综合运用的过程中可能缺乏坚实的理论基础，对改进结果无法提供强有力的支撑。

4 结 语

文本相似度方法研究已经取得诸多成果。国外学者首先提出文本相似度计算方法，国内学者基于国外研究进行了大量改进。本文纵观文本相似度计算方法的发展情况，对经典、新型算法进行了系统阐述和比较。通过分析，可以看出文本相似度的发展符合人类对事物的认知规律，经历了从感性到理性的过程。首先是字面方法，"看"上去相似则相似；然后以"词"为单位，采用词袋模型，上下文描述相似则相似；而文本中词语并不独立，词语之间的句法影响文本相似度判断，所以出现基于句法的方法；当已有方法仍存在语义不足问题时，研究人员则利用已积累的知识——本体，判断相似度；随着 Web3.0 的发展，网络资源成为不可忽视的宝贵财富，于是出现以网络知识为背景和基于搜索引擎的计算方法；神经网络算法的发展为文本表示带来新的灵感，出现词向量的文本表示方式。基于当前研究成果，笔者认为今后文本相似度计算方法的趋势有以下三个方面:

(1) 基于神经网络的方法研究将更加丰富。由于词向量表示文本，所表达的文本语义信息更符合人类认知，所以随着第三次人工智能浪潮的到来，神经网络算法将得到不断改进，基于神经网络的文本相似度计算也必将得到更多探索。

(2) 网络资源为文本相似度计算方法研究提供更多支持。Web3.0、移动网络以及未来 5G 技术的实现，网络资源无疑是最大、最丰富的语料库，与此同时语义网和关联数据进一步发展，网络文本资源面向结构

化与互连化。所以新型的信息组织结构与信息之间的链接方式将应用到文本相似度计算之中。

(3) 针对特定领域以及跨领域文本的相似度计算将成为今后发展的重点。跨学科合作越来越趋于常态化，领域专家的合作促进跨领域世界知识的集成并为跨领域文本的相似度计算提供便捷的人工参与和建议。

参考文献：

[1] Gomaa W H, Fahmy A A. A Survey of Text Similarity Approaches [J]. International Journal of Computer Applications, 2013, 68(13): 13-18.

[2] Pradhan N, Gyanchandani M, Wadhvani R. A Review on Text Similarity Technique Used in IR and Its Application [J]. International Journal of Computer Applications, 2015, 120(9): 29-34.

[3] 秦春秀，赵捧未，刘怀亮. 词语相似度计算研究[J]. 情报理论与实践，2007，30(1)：105-108. (Qin Chunxiu, Zhao Pengwei, Liu Huailiang. Research on Word Similarity Measurement [J]. Information Studies: Theory & Application, 2007, 30(1): 105-108.)

[4] 刘萍，陈烨. 词汇相似度研究进展综述 [J]. 现代图书情报技术，2012(7-8): 82-89. (Liu Ping, Chen Ye. Survey of the State of the Art in Word Similarity [J]. New Technology of Library and Information Service, 2012(7-8): 82-89.)

[5] 李慧. 词语相似度算法研究综述 [J]. 现代情报，2015, 35(4): 172-177. (Li Hui. A Review on the Research of Word Similarity Algorithms[J]. Journal of Modern Information, 2015, 35(4): 172-177.)

[6] 韩普，王东波，王子敏. 词汇相似度计算和相似词挖掘研究进展 [J]. 情报科学, 2016, 34(9): 161-165. (Han Pu, Wang Dongbo, Wang Zimin. Research Advancement in Word Similarity Calculation and Mining[J]. Information Science, 2016, 34(9): 161-165.)

[7] 孙海霞，钱庆，成颖. 基于本体的语义相似度计算方法研究综述 [J]. 现代图书情报技术，2010(1): 51-56. (Sun Haixia, Qian Qing, Cheng Ying. Review of Ontology-based Semantic Similarity Measuring[J]. New Technology of Library and Information Service, 2010(1): 51-56.)

[8] 刘宏哲，须德. 基于本体的语义相似度和相关度计算研究综述 [J]. 计算机科学, 2012, 39(2): 8-13. (Liu Hongzhe, Xu De. Ontology Based Semantic Similarity and Relatedness Measures Review[J]. Computer Science, 2012, 39(2): 8-13.)

[9] Lin D. An Information-theoretic Definition of Similarity [C]// Proceedings of the 15th International Conference on Machine Learning.1998.

[10] 刘群，李素建. 基于《知网》的词汇语义相似度计算 [J]. 中文计算语言学, 2002, 7(2): 59-76. (Liu Qun, Li Sujian. Word Similarity Computing Based on How-Net[J]. Chinese Computational Linguisties, 2002, 7(2): 59-76.)

[11] 董振东，董强. 知网[EB/OL]. [2016-12-08]. http://www.keenage.com/zhiwang/c_zhiwang.html. (Dong Zhendong, Dong Qiang. HowNet [EB/OL]. [2016-12-08]. http://www.keenage.com/zhiwang/c_zhiwang.html.)

[12] 田久乐，赵蔚. 基于同义词词林的词语相似度计算方法 [J]. 吉林大学学报：信息科学版，2010, 28(6): 602-608. (Tian Jiule, Zhao Wei. Words Similarity Algorithm Based on Tongyici Cilin in Semantic Web Adaptive Learning System [J]. Journal of Jilin University: Information Science Edition, 2010, 28(6): 602-608.)

[13] 张焕炯，王国胜，钟义信. 基于汉明距离的文本相似度计算 [J]. 计算机工程与应用，2001, 37(19): 21-22. (Zhang Huanjiong, Wang Guosheng, Zhong Yixin. Text Similarity Computing Based on Hamming Distance [J]. Computer Engineering and Applications, 2001, 37(19): 21-22.)

[14] Dice L R. Measures of the Aount of Ecologic Association Between Species [J]. Ecology, 1944, 26(3): 297-302.

[15] Harris Z S. Distributional Structure [A]// Papers in Structural and Transformational Linguistics[M]. Springer, Dordrecht, 1970.

[16] Salton G, Wong A, Yang C S. A Vector Space Model for Automatic Indexing [J]. Communications of the ACM, 1975, 18(11): 613-620.

[17] 郭庆琳，李艳梅，唐琦. 基于 VSM 的文本相似度计算的研究 [J]. 计算机应用研究，2008,25(11): 3256-3258. (Guo Qinglin, Li Yanmei, Tang Qi. Similarity Computing of Documents Based on VSM [J]. Application Research of Computers, 2008, 25(11): 3256-3258.)

[18] 李连，朱爱红，苏涛. 一种改进的基于向量空间文本相似度算法的研究与实现 [J]. 计算机应用与软件, 2012, 29(2): 282-284. (Li Lian, Zhu Aihong, Su Tao. Research and Implementation of An Improved VSM-based Text Similarity Algorithm [J]. Computer Applications and Software, 2012, 29(2): 282-284.)

[19] Landauer T K, Dumais S T. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge [J]. Psychological Review, 1997, 104(2): 211-240.

[20] Hofmann T. Probabilistic Latent Semantic Analysis [C]// Proceedings of the 15th Conference on Uncertainty in

Artificial Intelligence.1999.

[21] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[22] 王振振，何明，杜永萍. 基于 LDA 主题模型的文本相似度计算 [J]. 计算机科学，2013，40(12)：229-232. (Wang Zhenzhen, He Ming, Du Yongping. Text Similarity Computing Based on Topic Model LDA [J]. Computer Science, 2013, 40(12): 229-232.)

[23] 熊大平，王健，林鸿飞. 一种基于 LDA 的社区问答问句相似度计算方法 [J]. 中文信息学报，2012，26(5)：40-45. (Xiong Daping, Wang Jian, Lin Hongfei. An LDA-based Approach to Finding Similar Questions for Community Question Answer [J]. Journal of Chinese Information Processing, 2012, 26(5): 40-45.)

[24] 张超，陈利，李琼. 一种 PST_LDA 中文文本相似度计算方法 [J]. 计算机应用研究, 2016, 33(2): 375-377,383. (Zhang Chao, Chen Li, Li Qiong. Chinese Text Similarity Algorithm Based on PST_LDA [J]. Application Research of Computers, 2016, 33(2): 375-377,383.)

[25] Hinton G E. Learning Distributed Representations of Concepts[C]//Proceedings of the 8th Annual Conference of the Cognitive Science Society. 1986.

[26] Bengio Y, Ducharme R, Vincent P, et al. A Neural Probabilistic Language Model [J]. Journal of Machine Learning Research, 2003, 3(6): 1137-1155.

[27] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and Their Compositionality [C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. 2013.

[28] Pennington J, Socher R, Manning C D. GloVe: Global Vectors for Word Representation [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014: 1532-1543.

[29] Kenter T, Rijke M D. Short Text Similarity with Word Embeddings [C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. 2015: 1411-1420.

[30] Kusner M J, Sun Y, Kolkin N I, et al. From Word Embeddings to Document Distances [C]//Proceedings of the 32nd International Conference on Machine Learning. 2015.

[31] Huang G, Guo C, Kusner M J, et al. Supervised Word Mover's Distance [C]//Proceedings of the 30th Conference on Neural Information Processing Systems. 2016.

[32] Cilibrasi R L, Vitanyi P M B. The Google Similarity Distance [J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(3): 370-383.

[33] 刘胜久，李天瑞，贾真，等. 基于搜索引擎的相似度研究与应用 [J]. 计算机科学，2014，41(4)：211-214. (Liu Shengjiu, Li Tianrui, Jia Zhen, et al. Research and Application of Similarity Based on Search Engine [J]. Computer Science, 2014, 41(4): 211-214.)

[34] Sahami M, Heilman T D. A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets [C]//Proceedings of the 15th International Conference on World Wide Web. 2006: 377-386.

[35] 陈海燕. 基于搜索引擎的词汇语义相似度计算方法 [J]. 计算机科学, 2015, 42(1): 261-267. (Chen Haiyan. Measuring Semantic Similarity Between Words Using Web Search Engines [J]. Computer Science, 2015, 42(1): 261-267.)

[36] Hliaoutakis A. Semantic Similarity Measures in MeSH Ontology and Their Application to Information Retrieval on Medline [EB/OL]. [2016-12-08]. http://www.intelligence.tuc.gr/publications/Hliautakis.pdf.

[37] Batet M, Sanchez D, Valls A. An Ontology-based Measure to Compute Semantic Similarity in Biomedicine [J]. Journal of Biomedical Informatics, 2011, 44(1): 118-125.

[38] Rada R, Mili H, Bicknell E, et al. Development and Application of a Metric on Semantic Nets [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1989, 19(1): 17-30.

[39] Wu Z, Palmer M. Verb Semantic and Lexical Selection [C]//Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics. 1994:133-138.

[40] Richardson R, Smeaton A F, Murphy J. Using WordNet as a Knowledge Base for Measuring Semantic Similarity Between Words [EB/OL]. [2016-12-08]. http://pssd.computing.dcu.ie/wpapers/1994/1294.pdf.

[41] Li Y, Bandar Z A, McLean D. An Approach for Measuring Semantic Similarity Between Words Using Multiple Information Sources [J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(4): 871-882.

[42] Lin D. Principle-based Parsing without Overgeneration[C]//Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics. 1993.

[43] Resnik P. Semantic Similarity in a Taxonomy: An Information-Based Measure and Its Application to Problems of Ambiguity in Natural Language [J]. Journal of Artificial Intelligence Research, 1999, 11:95-130.

[44] Lord P W, Stevens R D, Brass A, et al. Investigating Semantic Similarity Measures across the Gene Ontology: The

Relationship Between Sequence and Annotation [J]. Bioinformatics, 2003, 19(10): 1275-1283.

[45] 边振兴. WordNet 中概念语义相似度 IC 参数模型研究 [J]. 计算机工程与应用, 2011, 47(19): 128-131. (Bian Zhenxing. Research on Model of IC Parameter for Semantic Similarity of Concept in WordNet [J]. Computer Engineering and Applications, 2011, 47(19): 128-131.)

[46] Tversky A. Features of Similarity [J]. Psychological Review, 1977, 84(4): 327-352.

[47] 葛斌, 李芳芳, 郭丝路, 等. 基于知网的词汇语义相似度计算方法研究 [J]. 计算机应用研究, 2010, 27(9): 3329-3333. (Ge Bin, Li Fangfang, Guo Silu, et al. Word's Semantic Similarity Computation Method Based on Hownet [J]. Application Research of Computers, 2010, 27(9): 3329-3333.)

[48] 王艳娜, 周子力, 何艳. WordNet中基于IC的概念语义相似度算法 [J]. 计算机工程, 2011, 37(22): 42-44. (Wang Yanna, Zhou Zili, He Yan. Concept Semantic Similarity Algorithm in WordNet Based on Information Content [J]. Computer Engineering, 2011, 37(22): 42-44.)

[49] 李文清, 孙新, 张常有, 等. 一种本体概念的语义相似度计算方法 [J]. 自动化学报, 2012, 38(2): 229-235. (Li Wenqing, Sun Xin, Zhang Changyou, et al. A Semantic Similarity Measure Between Ontological Concepts [J]. Acta Automatica Sinica, 2012, 38(2): 229-235.)

[50] 孙琛琛, 申德荣, 单菁, 等. WSR:一种基于维基百科结构信息的语义关联度计算算法 [J]. 计算机学报, 2012, 35(11): 2361-2370. (Sun Chenchen, Shen Derong, Shan Jing, et al. WSR: A Semantic Relatedness Measure Based on Wikipedia Structure [J]. Chinese Journal of Computers, 2012, 35(11): 2361-2370.)

[51] Strube M, Ponzetto S P. WikiRelate! Computing Semantic Relatedness Using Wikipedia [C]//Proceedings of the 21st National Conference on Artificial Intelligence. 2006.

[52] Gabrilovich E, Markovitch S. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis [C]// Proceedings of the 20th International Joint Conference on Artifical Intelligence.2007.

[53] Milne D, Witten I H. An Effective, Low-cost Measure of Semantic Relatedness Obtained from Wikipedia Links[C]// Proceedings of the 23rd Association for the Advancement of Artificial Intelligence. 2008.

[54] 盛志超, 陶晓鹏. 基于维基百科的语义相似度计算方法[J]. 计算机工程, 2011, 37(7): 193-195. (Sheng Zhichao, Tao Xiaopeng. Semantic Similarity Computing Method Based on Wikipedia [J]. Computer Engineering, 2011, 37(7): 193-195.)

[55] 彭丽针, 吴扬扬. 基于维基百科社区挖掘的词语语义相似度计算 [J]. 计算机科学, 2016, 43(4): 45-49. (Peng Lizhen, Wu Yangyang. Semantic Similarity Computing Based on Community Mining of Wikipedia [J]. Computer Science, 2016, 43(4): 45-49.)

[56] Lizorkin D, Medelyan O, Grineva M. Analysis of Community Structure in Wikipedia [C]//Proceedings of the 18th International Conference on World Wide Web. 2009: 1221-1222.

[57] 詹志建, 梁丽娜, 杨小平. 基于百度百科的词语相似度计算 [J]. 计算机科学, 2013, 40(6): 199-202. (Zhan Zhijian, Liang Li'na, Yang Xiaoping. Word Similarity Measurement Based on BaiduBaike [J]. Computer Science, 2013, 40(6): 199-202.)

[58] 尹坤, 尹红风, 杨燕, 等. 基于 SimRank 的百度百科词条语义相似度计算 [J]. 山东大学学报:工学版, 2014, 44(3): 29-35. (Yin Kun, Yin Hongfeng, Yang Yan, et al. Semantic Similarity Computation of Baidu Encyclopedia Entries Based on SimRank [J]. Journal of Shandong University:Engineering Science, 2014, 44(3): 29-35.)

[59] 穗志方, 俞士汶. 基于骨架依存树的语句相似度计算模型 [C]//1998 中文信息处理国际会议论文集. 1998. (Sui Zhifang, Yu Shiwen. The Skeletal-Dependency-Tree-Based Computational Model for the Sentence Similarity [C]// Proceedings of the International Conference on Chinese Computing.1998.)

[60] 李彬, 刘挺, 秦兵, 等. 基于语义依存的汉语句子相似度计算 [J]. 计算机应用研究, 2003, 20(12): 15-17. (Li Bin, Liu Ting, Qin Bing, et al. Chinese Sentence Similarity Computing Based on Semantic Dependency Relationship Analysis [J]. Application Research of Computers, 2003, 20(12): 15-17.)

[61] 李茹, 王智强, 李双红, 等. 基于框架语义分析的汉语句子相似度计算 [J]. 计算机研究与发展, 2013, 50(8): 1728-1736. (Li Ru, Wang Zhiqiang, Li Shuanghong, et al. Chinese Sentence Similarity Computing Based on Frame Semantic Parsing [J]. Journal of Computer Research and Development, 2013, 50(8): 1728-1736.)

[62] Blanco E, Moldovan D. A Semantic Logic-Based Approach to Determine Textual Similarity [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(4): 683-693.

[63] Jiang J J, Conrath D W. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy[C]// Proceedings of the

International Conference on Research in Computational Linguistics. 1997.

[64] Islam A, Inkpen D. Semantic Text Similarity Using Corpus-based Word Similarity and String Similarity [J]. ACM Transactions on Knowledge Discovery from Data, 2008, 2(2): 1-25.

[65] Tasi C S, Huang Y M, Liu C H, et al. Applying VSM and LCS to Develop an Integrated Text Retrieval Mechanism [J]. Expert Systems with Applications, 2012, 39(4): 3974-3982.

[66] 魏韡, 向阳, 陈千. 计算术语间语义相似度的混合方法 [J]. 计算机应用, 2010, 30(6): 1668-1670. (Wei Wei, Xiang Yang, Chen Qian. Combined Measurement Approach for Semantic Similarity of Terms [J]. Journal of Computer Applications, 2010, 30(6): 1668-1670.)

[67] Liu G, Wang R, Buckley J, et al. A WordNet-based Semantic Similarity Measure Enhanced by Internet-based Knowledge [C]//Proceedings of the International Conference on Software Engineering & Knowledge Engineering.2011.

[68] 王小林, 肖慧, 邰伟鹏. 基于 Hadoop 平台的文本相似度检测系统的研究 [J]. 计算机技术与发展, 2015, 25(8): 90-93. (Wang Xiaolin, Xiao Hui, Tai Weipeng. Research on Text Similarity Detection System Based on Hadoop [J]. Computer Technology and Development, 2015, 25(8): 90-93.)

[69] Atoum I, Otoom A. Efficient Hybrid Semantic Text Similarity Using Wordnet and a Corpus [J]. International Journal of Advanced Computer Science and Applications, 2016, 7(9): 124-130.

作者贡献声明:

陈二静: 提出研究选题, 起草及修改论文;
姜恩波: 修改论文。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail:chenerjing@mail.las.ac.cn。
[1] 陈二静.References.zip.参考文献.

收稿日期: 2017-05-09
收修改稿日期: 2017-06-15

Review of Studies on Text Similarity Measures

Chen Erjing[1,2] Jiang Enbo[1]
[1](Chengdu Documentation and Information Center, Chinese Academy of Sciences, Chengdu 610041, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: [**Objective**] This paper analyzes the popular text similarity measures and discusses their latest developments. [**Coverage**] We retrieved 69 key articles from CNKI and Web of Science databases by searching “TI: ‘text similarity’ or ‘semantic similarity’ or ‘lexical similarity’ ” in Chinese and English respectively. [**Methods**] We systematically reviewed the text similarity measures focusing on their basic concepts, characteristics and future directions. [**Results**] There were four types of text similarity measures: String-based, Corpus-based, Knowledge-based and others. Measures based on the neural network, Knowledge-based measures and inter-disciplinary measures could be the future research directions. [**Limitations**] We did not discuss the applications of those measures. [**Conclusions**] This paper is a comprehensive review of text similarity measure research.

Keywords: Text Similarity Semantic Similarity Ontology Bag of Words Model Neural Network